

DOCUMENT RESUME

ED 199 270

TM 810 169

AUTHOR Cantwell, Zita M.
TITLE Performance Tests as Dependent Variable Measures in Personnel Evaluation.
PUB DATE Oct 80
NOTE 29p.: Paper presented at the Annual Meeting of the Evaluation Network Conference (Memphis, TN, September 29-October 2, 1980).
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Performance Tests; *Personnel Evaluation; *Situational Tests; Test Bias; Test Reliability; Test Validity; *Work Sample Tests
IDENTIFIERS Generalizability Theory

ABSTRACT

Fair, valid, and reliable measurement of dependent variables is a concern in personnel evaluation studies. Performance tests, that is, job or work sample tests, situational tests, and job learning or trainability tests, are being used with increasing frequency as measurement techniques capable of meeting these needs. The three categories of performance tests are described with illustrations of their use. Issues related to development of fair, valid, and reliable performance tests are considered. Performance tests appear to be fair measures when individuals are classified by ethnic group, age, or sex. However, additional studies which combine task/environment samples with scoring procedures anchored to clearly specified aspects of job relevant behavior are needed. When handicapped individuals are to be evaluated, job learning or trainability tests appear to discriminate best compared to the other two categories of performance tests. Validity of performance tests is a function not only of content (including work environment) but also of scoring procedures. The behavioral consistency model can be applied to questions of validity arising from individual behavior changes which occur from training or entry to on-the-job. In some instances, generalizability coefficients provide the most useful information about the reliability of performance tests. (Author)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

Z. M. Cantwell

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Performance Tests as Dependent Variable

Measures in Personnel Evaluation¹

Zita M. Cantwell

Brooklyn College of the City University of New York

Introduction

As Nunnally and Wilson (1975) point out, valid, reliable, and fair measurement of dependent or criterion variables is a concern in personnel evaluation studies. A need exists for assessment procedures which yield information directly related to evaluation objectives which can be classified generally within five decision categories; (a) promotion, demotion, or layoff; (b) manpower planning and utilization; (c) salary adjustment; (d) communication between supervisors and subordinates; and (e) determination of management development needs. Traits or activities within these categories about which information is sought include (a) personal traits, e.g., cognitive and affective behaviors; (b) job behaviors; (c) managerial skills; and (d) achievement of previously established goals (Holley *et al.*, 1976).

Performance tests increasingly are being used to fill this need for data for personnel evaluation; a use for which they are well suited. A performance test is one in which, ideally, the criterion situation, that is, the job with its concomitant environment, is sampled and, except for direct on-the-job test situations, simulated with varying degrees of fidelity and comprehensiveness. The simulations can be identified on a continuum from work or job sample tasks in "off-the-job" settings to job learning and recognition tasks. Performance test format depends on the nature of the job task and does not exclude paper-

¹Based in part on a paper presented at the Evaluation Network Conference, Memphis, Tennessee, October 1980.

ED199270

TM 8/10/69

and-pencil measures (Anderson & Ball, 1975; Fitzpatrick & Morrison, 1971; Ryans & Frederikson, 1951; Siegel & Bergman, 1975). Performance tests are essentially narrow-band measures in terms of Cronbach and Gleser's (1965) definition of that construct. However, within categories, individual performance tests differ in the degree of generality of application; differences chiefly related to distance of content from the real job/environment situation.

A number of studies have shown that performance tests appear to yield information which can meet EEOC guidelines; offer fair estimates irrespective of ethnic, cultural, sex, and age group membership; predict trainability better than general ability measures; are appropriate for evaluating continuing competency of practicing professionals; and, are useful and fair bases for evaluating performances of handicapped workers or job applicants (Brugnole et al., 1979; Bunker & Cohen, 1977; Gordon & Kleinman, 1976; Hamner et al., 1974; Kelleher & Quirk, 1976; McGuire, 1979; Muchinsky, 1975; Stodden et al., 1977). However, in addition to showing the strength of performance tests as dependent variable measures in personnel evaluation, these studies pinpoint issues related to their development and use which are in need of further study, for example, issues in the areas of test fairness, scoring, validity, and reliability. This paper discusses the use of performance tests in personnel evaluation with an emphasis on specific aspects of these issues.

Using Gordon and Cohen's (1973) distinction, literature selected for review is focused on assessments of individual performance--evaluations of concrete, observable behaviors or products of behavior--rather than on assessments of effectiveness--evaluations of functioning of organizations or administrative units using evaluations of performances of groups of individuals. Literature on evaluation of teacher performance has not been included.

Categories of Performance Tests

McClelland (1973) has stated that the best test is one based on criterion (performance) sampling. Performance tests are based on a general assumption that measuring the functioning of an individual under testing conditions which transfer directly to actual job behavior can generate the most authentic data on current functioning status. For example, Hinrichs (1969), in a study comparing "real life" assessments of management potential with performance tests, paper-and-pencil tests, and personality inventories, found that the closer the measurement device came to reality the better it was as a predictor. The systematic observation of behavior in the natural occupational environment can provide the truest sample of functioning. However, the further the sampled job activities are lifted out of the actual task context, the more liable are the results to serious distortion (Bersoff, 1973; Guion, 1965).

All performance tests can be described and compared with reference to four essential, continuous, and interdependent dimensions: (a) directness--reality level--of tasks and situations, (b) type of criteria, (c) standardization conditions, and (d) objectivity of scoring (Bersoff, 1973; Osborn, 1977; Rampton, 1978). With respect to the first dimension, performance tests usually can be subsumed under one of three categories: work sample, situational, or rehearsed performance tests.

Work sample tests are "controlled" tryouts under actual conditions of the working situation. They range from performances measured directly on the job to performances measured in off-the-job settings in the work place. Performance test tasks are selected to have point-to-point correspondence with the criteria. Tasks can be distinguished by type between those having clearly specified correct and incorrect responses and those whose scores are based on judgments of observers (Asher & Sciarrino, 1975; Guion, 1965; Ryans &

Frederikson, 1951). The criterion for task selection is correspondence to work task/environment (Boyd, 1970; Fitzpatrick & Morrison, 1971). In this context, the distinction between activity task and paper-and-pencil task is irrelevant--performance tests in accounting are usually paper-and-pencil (Garrison, 1979).

Applied performance tests, subsets of work sample tests, are those in which the stimulus or response or both in a test task mirror real life situations (Sanders, 1976; Sanders & Sachse, 1977). Applied performance tests have been used most frequently to evaluate performance after training.

Situational tests are described by Frederikson (1975) as tests in which the tasks and environments presented approach, in varying degrees, the real work situation. All testees are given identical problems and told to behave "as if" they were actually in a real life situation. Situational tests fall between work sample and on-the-job performance tests and standardized group tests of aptitudes, abilities, personality traits. They are an attempt to measure more complex qualities than isolated aptitudes, traits, and skills, and at the same time, to retain more control over the testing situation than is possible in on-the-job performance tests or work sample tests (Fiske, 1954). The in-basket test (Gall, 1979; Levine & Flory, 1973); TOWER, WREST, JEVS, Singer Vocational Evaluation Systems (Backman et al., 1977; Stodden et al., 1977); symbolic troubleshooting tests (Gustafson, 1976); computer-based simulation (Newsom et al., 1978); and, synthetic performance tests (Osborn & Ford, 1976) are examples of situational tests.

Situational tests differ in the extent to which the complexity of behaviors assessed replicate "real life" situations. For example, synthetic performance tests are designed to focus on assessment of (a) task condition or (b) task behavior in a simulated situation. Another situational test model presents a more complex design in which tasks yield information to permit as-

assessment on three dimensions: (a) one of three sets of behavior--cognitive, perceptual, or psychomotor; (b) one of two skills levels--high (task needs practice) or low (no practice needed); and (c) with or without a motivational or affective component (Osborn & Ford, 1976; Olmstead, 1974).

Situational tests can be more difficult to score than other types of performance tests. Point-to-point correspondence between the job/task environment and the test task/environment is not as clearly established in the simulated testing situation. Therefore, descriptions of criterion behaviors bend more toward the general, rather than the specific job, case. More than in the case of work sample or on-the-job tests, situational tests call for observer judgment in arriving at scores. For example, Kelly (1954) found situational testing to be only marginally effective in evaluating clinical psychologists because of difficulties encountered in scoring the tests for the criterion behaviors.

Rehearsed performance test is a phrase sometimes used to categorize job learning tests (Siegel et al., 1977; Siegel & Bergman, 1975, 1972; Siegel, Bergman, & Lambert, 1978, 1973; Stodden et al., 1977) and trainability tests (Downs, 1970; Robertson & Downs, 1979; Smith & Downs, 1975). Such assessments combine an instruction period with a subsequent performance test to determine what has been learned during that period. The assessment is based on the rationale that if the task to be learned replicates critical components of the job task, estimates can be made of the likelihood of success in learning to do the job. The task to be learned must (a) be based on critical elements of the job, (b) use only such skill and knowledge as can be covered during the learning period, (c) be sufficiently complex to allow a range of errors, and (d) be able to be completed within a reasonable time frame. The potential trainee

is an active participant in the learning situation. A checklist based on probable errors is used to evaluate performance on the test. The possibility of incorporating training for evaluation tasks has been suggested in conjunction with the in-basket test and assessment center tasks in general. However, validation and cross-validation of predictors appear to present problems needing solution before predictive efficiency of learning performance tests can be estimated (Cohen & Penner, 1976).

The type of performance test selected or constructed depends on (a) the reasons for the personnel evaluation; (b) the use to be made of the device; and (c) the objectives for which it can supply information, predict job success, diagnose performance deficiencies, evaluate training, and provide criterion measures for promotion or advancement and research (Guion, 1965; Ryans & Frederikson, 1951). However, there is not clear agreement on the categories of behaviors that the different types of performance tests are able to measure. Muchinsky (1975) holds that the utility of the work sample test as a selection device lies in part in its potential ability to manifest a high degree of content validity, reliability, and practicality. Yet, he and others (for example, Ronan *et al.*, 1976) question whether it is possible to measure some performance dimensions, e.g., interpersonal relations or motivation, by work sample tests. Other research (for example, Gordon, 1967) has demonstrated that performance tests can yield valid and reliable estimates of affective behaviors.

Compared to tests in general, performance tests are constructed to be narrow band, high fidelity instruments. However, band width varies as the test tasks/environment move from exact point-to-point correspondence with the criterion. Selection or construction decisions need to take account of (a) the manner in which criterion behaviors are to be defined and sampled; for example, with reference to objectives stated behaviorally, "amplified" objectives,

or domain specifications (Hambleton et al., 1978); (b) whether the product or process is to be measured (Osborn, 1974); and, (c) content issues of situational--job and environment--specificity and personal trait specificity (Kavanaugh, 1971). Perhaps, the most difficult problem to handle psychometrically is that of control over the testing situation and contents. The more tenuous the control, the greater the opportunity for bias in the job sample (Blood, 1974). The enigma: control becomes less as the performance test moves closer to on-the-job assessment, yet this closeness is the essential quality of the performance test.

Performance tests are being used successfully in personnel evaluation. For example, Cascio and Phillips (1979) report data on a series of 21 performance tests (11 motor and 10 verbal) used to evaluate 263 applicants, entry level or promotional, for city government jobs in one city over a 17-month period. Testers and applicants were male and female; Black, white, and Spanish speaking. No adverse impact was observed. Significant reduction (to zero) in the number of complaints about hiring and promotion testing practices was among the advantages stated. The program used two of the three general categories of performance tests: work sample tests and job learning tests. The U. S. Army has implemented a combined verbal/hands-on performance certification testing program--the Skill Qualification Test (SQT) battery (Hirshfeld, 1976; Wiltshire, 1978). The in-basket exercise, an example of a generalizable performance test, receives wide use especially in assessment of managerial or supervisory potential (Gill, 1979).

What is the nature of the balance sheet for performance tests on specific issues in need of further study: fairness, general issues of validity, behavioral consistency model, and reliability--the theory of generalizability.

Fairness

Doppelt and Bennett (1967) note that the charge of discrimination is brought against testing most frequently on the grounds that a test (a) lacks relevance or (b) is unfair in content for the individual being assessed. Fairness in a performance test has two additional dimensions: (a) fairness in scoring and (b) control of irrelevant sources of variance which otherwise interact with the performance being measured (Guion, 1978).

O'Leary (1973) writes that the more nearly the test content duplicates the specific job tasks/environment the greater are the chances of developing fair selection devices. The characteristics of the specific task to be sampled are those of the job and its environment, not those of the worker doing the job. The job exists independently of the worker.

Test fairness is not a simple issue to study. From a theoretical perspective, bias is both a statistical and a value oriented construct. There are approximately five definitions of statistical bias (Cleary, 1968; Cole, 1973; Darlington, 1971; Linn & Werts, 1971; Thorndike, 1971). And as Linn, Hunter, and Schmidt (Linn, 1973; Hunter, J. E. & Schmidt, 1976; Schmidt & Hunter, 1974) note, while each of the definitions is statistically reasonable all are in conflict. Linn suggests that the resolution of the problem of fairness extends beyond the technical solution and depends ultimately on weighing competing values attached to a decision. What have studies completed shown about fairness with respect to race, age, handicap, and sex of performance tests used for personnel evaluation?

Studies of racial bias in performance tests have yielded results which can be of use in test development and selection. Two studies considered in juxtaposition identify critical factors in bias. Hamner et al. (1974) completed a "lab." study using college students of mixed sex and ethnic groups

as raters and ratees. Raters evaluated a task performance, a filmed performance of potential employees stocking grocery shelves, using a global rating scale. Sex and race stereotypes appeared to influence the results: high ratings were given to Ss of the same race, higher ratings were given to Black Ss who did poorly than to white Ss who did poorly, higher ratings were given to female Ss (the job was unskilled and traditionally male dominated). In the second study, Brugnot et al. (1979) had 56 maintenance mechanics evaluate videotapes of a Black applicant and a white applicant performing (a) a relevant task and (b) an irrelevant task. Evaluation was completed using a highly specific behavioral rating form, a global rating scale, or both. Race linked bias was found only when Ss (the raters) were asked to make global evaluation of an irrelevant task. The investigators concluded that careful work sample development procedures combined with rater focus on carefully specified aspects of relevant behavior can make the potential for racial bias in performance tests small.

Gael et al. (1975), using Cleary's definition of bias, evaluated 10 paper-and-pencil employment tests for white, Black, and Spanish speaking job candidates against a Clerical Performance Maintenance (CPM) package. No bias was found. The investigators noted they now had two batteries of valid tests for use in personnel evaluation, but were surprised at how difficult the development of the performance (CPM) battery was. Schmidt and his colleagues (Schmidt et al., 1973, 1974, 1977; Schmidt & Hunter, 1974) found no differential validity for ethnic groups in performance tests. D. R. Hunter (1976) and Feldt et al. (1977) found performance tests discriminated better than paper-and-pencil tests among mixed ethnic groups at different plant locations. Siegel, Bergman, and Lambert (Siegel & Bergman, 1972; Siegel, Bergman, & Lambert, 1978) found no racial bias in job learning tests.

A six year study by Campbell and his colleagues (Campbell et al., 1975) of selection and promotion across racial groups for jobs of medical technician, cartographic technician, and inventory management specialist found criterion performance measures valid for one ethnic group were valid for all. The measures predicted equally well for majority and minority groups. Several researchers have noted that performance tests are perceived by majority and minority examinees as fairer, clearer, and more appropriate in difficulty level than paper-and-pencil tests (Cascio & Phillips, 1979; Schmidt et al., 1977). However, test bias is not necessarily present if significant differences exist between majority and minority groups because the same factors may lead to low performance on the criterion and on the predictor (Guion, 1966).

Farr et al. (1973) hypothesized that use of a work sample preemployment test for a sample of sewing machine operators from different ethnic backgrounds would lead individual operators to more accurate expectations about task requirements and result in higher job refusal rate and lower voluntary turnover rate. The hypothesis was supported for white but not Black Ss. Job expectations appeared to differ among ethnic groups and these expectations affected individual response orientations to work sample tests. Interaction of motivational and environmental variables in performance test scores has received only limited attention in the research.

Performance tests appear particularly suitable for a broad spectrum of different ages. For example, speed and test wiseness factors, which could bias results in favor of younger testees, are more easily removed from or controlled in performance tests. Initial impetus for development of trainability tests was in part the need to retrain older workers to meet technological changes in industry combined with lack of suitable assessment devices to use with older workers. Training tests have been successful in meeting the need

in occupational situations for which they have been developed: chiefly, non-managerial and non-supervisory jobs (Downs, 1970). Kuriansky et al. (1976) found a simple performance test more accurately assessed self-care capacities of hospitalized geriatric psychotic patients than did patients' self-reports.

However, in the area of age bias and performance tests more research is needed. Many general studies of attitudes toward older workers (Kirchner & Dunnette, 1954; Kelleher & Quirk, 1976) indicate that a relationship exists between age of evaluator and evaluatee and score given by the former. Schwab and Heneman's (1978) study of age stereotyping in performance appraisal appears to reinforce this finding. In that study, older raters gave lower ratings to older Ss; younger raters gave higher ratings to older Ss.

Stodden et al. (1977) point to bias in performance tests now being used with mentally retarded workers and job applicants and suggest the need for research to develop job learning tests or trainability tests which meet the needs of these individuals. The tests should include both the work and the motivational/interest aspects of job performance. A study by Finkle (1975) found that blind individuals performed significantly better on job learning and trainability tests, that is, on a performance test preceded by a series of simulated training sessions.

Sex bias as well as racial bias was found by Hamner et al. (1974) in their "lab." study of the evaluation of performance of an unskilled job. A global rating scale was used to assess task performance. Hamner et al.'s findings were not replicated when behaviorally anchored rating scales were used to evaluate a relevant task. Huck (1975) found assessment center protocols, which included an in-basket test, equally predictive for white and Black females. Paper-and-pencil tests added little or nothing to the predictive efficiency of the battery. Schmitt et al. (1978) studied pairs of

validity coefficients from studies reporting validities for males and females which were published in Educational and Psychological Measurement, Journal of Applied Psychology, and Personnel Journal from 1955 to 1978--a total of 6,219 coefficients, including some from performance tests. Women appeared to be slightly, but significantly, more predictable than men. As with age and handicap, additional research is needed on the interactions of this variable with aspects of performance tests.

Validity

Validity statements relate to the reasonableness of interpretations of test performances (Cronbach, 1980). But as Ebel (1977, 1961) points out, the concept of validity in general is not as clearly prescribed in practice as definitions would appear to indicate. Personnel decisions not infrequently necessitate trade-offs between validity and dependability (Bunch, 1978). Not only must the content--job task/job environment samples or simulations--of performance tests be deemed valid, but the procedures for scoring the performance tests also must possess validity (Brass & Oldham, 1976). Further, interpretations of results of validity studies of performance tests is challenging because, as recent studies have emphasized, sample sizes used are frequently too small to produce adequate power to support validity statements (Cascio et al., 1978; Schmidt et al., 1976).

Studies of performance test validity can be subsumed under three categories: (a) point-to-point theory, (b) content versus construct validity, and (c) concurrent and predictive validity.

Point-to-point theory. Most performance tests have the opportunity to exhibit point-to-point correspondence with the criterion, namely, the job and its environment, and therefore have been said to possess a high degree of validity; at least, of face validity (Asher & Sciarrino, 1974; Bersoff, 1973; Inskeep,

1971). And overall, studies have found performance tests closely related to job tasks to be very effective evaluators of job performance (Gordon, 1967; Gordon & Kleiman, 1976; Hunter, D. R. & Thompson, 1978; Hunter et al., 1977; McLaurin, 1974; Schmidt et al., 1974, 1973; Talbert et al., 1976; Thornton, 1971). Yet, assessing validity of performance tests is complicated additionally because of the difficulty in causal attribution in work performance changes; this difficulty becomes more acute as the performance task becomes more realistic and as selection bias in job samples cuts into the point-to-point relationship (Blood, 1974; Cook & Campbell, 1976). Further, point-to-point correspondence analyses are made most frequently between the job tasks and the tasks sampled for the test. They rarely are designed to compare the job task in the performance environment with the sampled job task/environment specifics of the performance test. Nor is scoring of performances usually considered in these analyses. The need to sample the entire job, that is, task plus environment, and concomitant scoring procedures is recognized but not much studied.

Content versus construct validity. Content validity, in the sense of defining content from a construct, is sometimes described as a special case of construct validity (Guion, 1974). While Ebel (1977) holds that confusion would be lessened by a return to the original "latent trait" definition of construct, Cronbach (1980) has stated that the definition should be expanded to include environments as well as traits. Both Ebel (1977) and Guion (1978, 1974) suggest that rather than distort classical concepts of validity it is in most cases adequate--and will reduce confusion--to refer to and describe the representativeness of the content sampling of performance tests.

However, as Guion (1978) continues, it frequently may be necessary to show formally that job related content is scored fairly. Depending on the

design and objective of the performance test, either the product or the process of the activity may be scored. In either event, evidence must be presented that the scoring is valid. Borman (1978) states that even under "ideal" conditions, variance among raters can cause validity problems with performance rating because the raters (a) do not have time to observe all the behaviors which can be rated, (b) are inexperienced in performance evaluation, (c) are working with a format not well related to the job, or (d) are operating under organizational constraints. Greenwood & McNamara (1967) found interrater agreement "reasonably" high. Several other investigators have discussed the problem of unbiased scoring but more research is needed on the topic. Guion (1978) suggests disconfirmity studies investigating specifically hypothesized constraints and making use of multivariate techniques of analysis be completed to determine whether bias exists in scoring procedures.

Representativeness of content sampling, as noted previously, is the single most important issue in content validity (Guion, 1978; Wollach, 1976). Job analysis becomes the backbone of content validation. A multimethod approach, which samples the task requirements, characteristics of the task setting, and the abilities and motives of the person being evaluated has been used in some studies (Prien, 1977; Prien & Ronan, 1971). Another technique, multidimensional scaling, can be adapted to describe the job and its setting (Rampton, 1978; Siegel & Smith, 1976). McCormick and his associates (1972; 1957) have developed and tested out in practice the Positional Analysis Questionnaire (PAQ), designed to identify systematically job characteristics and dimensions.

Abilities measured by work sample tests are extracted by application of construct validation techniques (Bachman et al., 1977; Jones & Whittaker, 1973). However, Jones and Whittaker's (1973) study, which looked at the same work

sample test applied to a group of trainees and a group of workers, similar in selected characteristics to the trainees, who had completed the same training program two years earlier, found the test measuring different abilities in the two groups.

Concurrent and predictive validity. Concurrent validity data for work sample tests are held to give no better information on the predictive power of the tests than no validity data at all since substituting concurrent for predictive validity using worker performance data ignores the effects of motivation and job expertise on validity (Guion, 1965; Hinrichs, 1969). Concurrent validity, on the other hand, is important for the development of performance test scoring scales and for identifying criterion measures in studies of predictive validity (Guion, 1965; Meyer, 1970).

A wide variety of studies shows that predictive validity tends to be higher for performance tests compared to other measures, e.g., paper-and-pencil tests or supervisory ratings, used to evaluate job performance or predict success in training, in jobs ranging from managerial to semi-skilled and using performance tests from three categories: work/job sample, situational, and job learning or trainability tests (Cohen & Penner, 1976; Gael, 1975; Gordon, 1967; Gordon & Kleiman, 1976; Inskip, 1971; McLaurin, 1974; Talbert et al., 1976; Thorndike, 1971).

Nonetheless, predictive validity studies of performance tests must overcome the problems inherent in the fact that different abilities are measured by a given test during training and on the job (Asher & Sciarrino, 1974; Ebel, 1977; Guion, 1978; Jones & Whittaker, 1973). As Bender and Loveless (1958) note in a report on a series of evaluation studies of successive training classes, validity coefficients which exist over a two year span may not have the same magnitude over a longer period, say five years. The possibility

stands that applicants evaluated and rejected by the short-range criteria would have, if hired, performed the required task better for a longer period of time than those evaluated and employed.

Studies are beginning to take into account the differences in job and work setting expectations among ethnic subgroups (Farr et al., 1973). Findings from this research could affect the relationship of the worker to the job in ways not accounted for in current predictive validity studies.

Wernimont and Campbell's (1968) behavioral consistency model presents a framework for study of many of the issues raised about performance test validity which concern changes in behavior over time.

Behavioral Consistency Model

The emphasis on the need to establish consistencies between relevant dimensions of behaviors, e.g., job behaviors and preemployment behaviors, which appeared in Wernimont and Campbell's (1968) article, gave an impetus to the use of performance tests. The model emphasizes a focus on samples of actual behaviors needed for a job as descriptors of behavior, rather than on test scores as signs of possible performance, in personnel evaluation studies. Actual job behaviors are analyzed within a framework of consistency. As Campion (1972) notes, the model has been too little studied for the promise it shows. Yet there findings on issues related to behavioral consistency which can affect the use of performance tests as dependent variable measures.

Studies of a wide range of jobs from managerial through unskilled which have been designed within the behavioral consistency framework have found performance tests consistently more accurate predictors of job function than paper-and-pencil tests (Hinrichs, 1969; Mount et al., 1977). An unanswered research question related to this predictive strength remains: Why do work samples possess validity--because of common elements between the performance

and job task or because of worker interest and motivation?

Ability requirements for the task performance change over a period of training. Some research shows systematic changes in abilities required to perform certain tasks over time which indicate that general ability measures are more adequate for evaluation at certain periods of training, performance tests at others. Other studies find general ability tests of little value at any period in training (Fleishman, 1967, 1957; Hinrichs, 1970). These findings lead into the question of temporal stability of performance. Schmidt et al. (1974) and Shipley (1978) emphasize the need for data on this issue as it relates to performance tests.

Reliability

Just as with validity, it is not possible to estimate the reliability of a performance test until the reliability of judging performance has been shown to be adequate (Fitzpatrick & Morrison, 1971). Given this fact, methods of estimating reliability of performance tests are open to discussion. The challenge rests in the generalizability or appropriateness of the job behaviors and environmental conditions sampled (Ebel, 1977). For, as Nunnally and Wilson (1975) note, the major criteria by which strategies and measurement techniques for evaluating human behaviors must be judged is the extent to which generalization is promoted or hindered. Yet, a quality aimed for in performance test design is point-to-point correspondence with the criterion which, if realized, must narrow generalizability.

Suggestions offered for improving reliability coefficients obtained for performance tests emphasize the need for specificity of criterion behaviors and the appropriateness of criterion behaviors for the task and environment (Bourgeois & Slivinski, 1974; Greenwood & McNamara, 1967; O'Connor, 1974). Use of synthetic performance tests has been suggested to increase control over

criterion behaviors (Fiks, 1976; Skakun, 1979). However, changes in behavior over time are due to various factors over which the psychometrist may be able to exercise little control. For example, work methods alter (Corlett et al., 1971; Salvendy, 1975; Steel et al., 1945); practice effects are present differently in performance, depending to a great extent--in studies evaluating performance after training compared to after being on the job--to the specific skills used on the job (Pederson, 1978). Further, behaviors contributing to job performance change over time in unpredictable ways (Fleishman, 1967; Schmidt et al., 1974; Shipley, 1978).

One approach to looking at reliability of performance tests is to work with coefficients of generalizability rather than classical reliability coefficients (Cronbach et al., 1972; 1963).

Theory of generalizability. Fidelity refers to the relationship between test conditions and real conditions and depends on how well test conditions transfer to real world behavior (Borman et al., 1978). As Fitzpatrick and Morrison (1971) state, a valid performance test is one with complete fidelity and comprehensiveness. One way of estimating reliability of a performance test is to systematically assess effects on performance as test content moves from reality. Cronbach, Glesser, et al.'s (1972; 1963) theory of generalizability can be applied to performance test data to obtain an index of effects of movement away from reality. Different from classical reliability theory, generalizability theory requires the researcher to constrain the universe to which a response may be generalized, that is, to specify the universe of behaviors which is of interest to him, before he can begin to study generalizability. The universe is described unambiguously; the conditions which fall within it are clear. Classical theory, which assumes that a test has one true score and one reliability coefficient, differs from generalizability theory. The latter

assumes scores and coefficients specific to situations.

The theory has been in use, but not widely, in performance testing. Morse and Morse (1976) provide a model based on the theory which has potential for use with performance tests. As Skaken et al. (1974), Sanders (1976), and others note, there is need for research on the issues of generalizability especially when working with synthetic performance tests.

Conclusion

Performance tests are reasonable choices for use as dependent variable measures in personnel evaluation. In selected situations, they have proved to be effective in meeting EEOC guidelines and reducing ethnic, age, sex, and handicap bias in assessment. They usually are considered "fair" by testees. Studies support them as valid and, depending on the type of performance test, generally reliable. However, performance tests present continuing challenges related to their development and improvement.

Among issues in need of further study are those related to (a) fairness of performance tests among specific subgroups focusing especially on scoring techniques, (b) development of job learning or trainability tests for handicapped workers and job applicants, (c) validation studies on scoring procedures, (d) categories of behaviors which can be measured validly by performance tests--by type of test, (e) applications of generalizability theory as an index of fidelity of performance tests as content departs from specific representation of reality, (f) construction of performance tests based on unbiased sampling of work tasks combined with environmental conditions, (g) application of behavioral consistency model to questions of performance test validity arising from individual behavior changes from training or entry to on-the-job, (h) situation specific characteristics of performance tests and the development of more generalizable instruments, and (i) interaction of

personal characteristics, e.g., motivation and environmental variables, with performance test scores.

References

- Anderson, S. B., Ball, S., Murphy, R. T., & Associates. Encyclopedia of Educational Research. San Francisco: Jossey-Bass, 1975.
- Asher, J. J., & Sciarino, J. A. Realistic work sample tests: A review. Personnel Psychology, 1974, 27, 519-533.
- Backman, M. E. et al. Examination of the Aptitudes Measured by Work Samples in MICRO-TOWER Evaluation System. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York City, 1977. (ERIC Document Reproduction Service No. ED 152 784)
- Bender, W. R. G., & Loveless, H. E. Validation studies involving successive classes of trainee stenographers. Personnel Psychology, 1958, 11, 491-508.
- Bersoff, D. N. Silk purses into sow's ears. The decline of psychological testing and a suggestion for its redemption. American Psychologist, 1973, 28, 892-899.
- Blood, M. R. Job samples: A better approach to selection testing? American Psychologist, 1974, 29, 218-219.
- Borman, W. C. Exploring upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 1978, 63, 135-144.
- Bourgeois, R. P., & Slivinski, L. W. The inter-rater reliability of the Consolidated Fund In-Basket. Studies in Personnel Psychology, 1974, 6, 47-51.
- Boyd, J. L., & Shimberg, B. Handbook of Performance Testing. Princeton, N.J.: Educational Testing Service, 1979. (ERIC Document Reproduction Service No. ED 052 220)
- Brass, D. J., & Oldham, G. R. Validating an in-basket test using an alternate set of leadership scoring dimensions. Journal of Applied Psychology, 1976, 61, 652-657.
- Brugnot, G. A., Campion, J. E., & Basen, J. A. Racial bias in the use of work samples for personnel selection. Journal of Applied Psychology, 1979, 64, 119-123.
- Bunch, M. B. Making Decisions About Adult Learners Based on Performance on Functional Competency Measures. Paper presented at the Annual Adult Education Research Conference, San Antonio, 1978. (ERIC Document Reproduction Service No. ED 154 137)
- Bunker, K. A., & Cohen, S. L. The rigors of training evaluation: A discussion and field demonstration. Personnel Psychology, 1977, 30, 525-541.

- Campbell, J. T. et al. An Investigation of Sources of Bias in the Prediction of Job Performance: A Six-Year Study. New York: Ford Foundation, 1975. (ERIC Document Reproduction Service No. ED 109 155)
- Campion, J. E. Work sampling for personnel selection. Journal of Applied Psychology, 1972, 56, 40-44.
- Cascio, W. F., & Phillips, N. F. Performance testing: A rose among thorns. Personnel Psychology, 1979, 32, 751-766.
- Cascio, W. F., Valenzi, E. R., & Selbey, V. Validation and statistical power: Implications for applied research. Journal of Applied Psychology, 1978, 63, 586-595.
- Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.
- Cohen, S. L., & Penner, L. A. The rigors of predictive validation: Some comments on "A job learning approach to performance prediction." Personnel Psychology, 1976, 29, 595-600.
- Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.
- Cook, T. D., & Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In Dunnette, M. D. (Ed.) Handbook of Industrial and Organizational Psychology. Chicago: Rand-McNally, 1976. Pp. 223-326.
- Corlett, E. N., Salvendy, G., & Seymour, W. D. Selecting operators for fine manual tasks: A study of the O'Connor Finger Dexterity Test and Perdue Peg-board. Occupational Psychology, 1971, 45, 57-65.
- Cronbach, L. J. Selection theory for a political world. Public Personnel Management, 1980, 9, 37-50.
- Cronbach, L. J., & Gleser, G. C. Psychological Tests and Personnel Decisions (2nd Ed.) Urbana: Univ. Illinois Press, 1965.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley, 1972.
- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 71-82.
- Darlington, R. D. Another look at "culture fairness." Journal of Educational Measurement, 1971, 8, 71-82.
- Doppelt, J. P., & Bennett, G. K. Testing job applicants from disadvantaged groups. Test Service Bulletin No. 57. New York: Psychological Corp., 1967.

- Downs, S. Predicting training potential. Personnel Management, 1970, 2(9), 26-28.
- Ebel, R. L. Comments on some problems of employment testing. Personnel Psychology, 1977, 30, 55-63.
- Ebel, R. L. Must all tests be valid? American Psychologist, 1961, 16, 640-647.
- Farr, J. L., O'Leary, B. S., & Bartlett, C. J. Effect of work sample test upon self-selection and turnover of job applicants. Journal of Applied Psychology, 1973, 58, 283-285.
- Feild, H. S., Bayley, G. A., & Bayley, S. M. Employment test validation for minority and nonminority production workers. Personnel Psychology, 1977, 30, 37-46.
- Fiks, A. I. et al. Public Assistance Worker Job Trial. Washington, D.C.: Department of Labor, Employment & Training Administration, 1976. (ERIC Document Reproduction Service No. ED 133 463)
- Finkle, L. J. The effect of simulated testing sessions on work-sample performance scores of blind individuals (Doctoral dissertation, George Peabody College for Teachers, 1974). Dissertation Abstracts International, 1975, 35, 4283A. (University Microfilms No. 74-29, 166)
- Fiske, D. W. Why do we use situational performance tests? Personnel Psychology, 1954, 7, 464-469.
- Fitzpatrick, R., & Morrison, E. J. Performance and product evaluation. In Thorndike, R. L. (Ed.) Educational Measurement (2nd Ed.) Washington, D.C.: American Council on Education, 1971. Pp. 237-270.
- Fleishman, E. A. Individual differences and motor learning. In Gagné, R. M. (Ed.) Learning and Individual Differences. Columbus, O.: Chas. Merrill, 1967. Pp. 165-191.
- Fleishman, E. A. A comparative study of aptitude patterns in unskilled and skilled psychomotor performance. Journal of Applied Psychology, 1957, 41, 263-273.
- Frederikson, N. Situational tests. In Anderson, S. B. et al. Encyclopedia of Educational Evaluation. San Francisco: Jossey-Bass, 1975. Pp. 371-374.
- Gael, S., Grant, D. L., & Ritchie, R. J. Employment test validation for minority and nonminority clerks with work sample criteria. Journal of Applied Psychology, 1975, 60, 420-426.
- Garrison, L. L. Objective performance testing in accounting. Journal of Business Education, 1979, 54, 313-316.
- Gill, R. W. T. The in-tray (in-basket) exercise as a measure of management potential. Journal of Occupational Psychology, 1979, 52, 185-197.

- Gordon, L. V. Clinical, psychometric, and work-sample approaches in predicting success in Peace Corps training. Journal of Applied Psychology, 1967, 51, 111-119.
- Gordon, M. E., & Cohen, S. L. Training behavior as a predictor of trainability. Personnel Psychology, 1973, 26, 261-272.
- Gordon, M. E., & Kleiman, L. S. The prediction of trainability using a work sample test and an aptitude test: A direct comparison. Personnel Psychology, 1976, 29, 243-253.
- Greenwood, J. M., & McNamara, W. J. Interrater reliability in situational tests. Journal of Applied Psychology, 1967, 51, 101-106.
- Guion, R. M. Scoring of content domain samples: The problem of fairness. Journal of Applied Psychology, 1978, 63, 499-506.
- Guion, R. M. Open a new window: Validities and values in psychological measurement. American Psychologist, 1974, 29, 287-296.
- Guion, R. M. Employment tests and discriminatory hiring. Industrial Relations, 1966, 5, 20-37.
- Guion, R. M. Personnel Testing. New York: McGraw-Hill, 1965.
- Gustafson, H. W. Symbolic troubleshooting tests: A description and methodological review. Catalog of Selected Documents in Psychology, 1976, 6, 81.
- Hambleton, R. K. et al. Criterion-referenced testing and measurement: A review of technical developments. Review of Educational Research, 1978, 48, 1-47.
- Hamner, C. W., Kim, J. S., Baird, L., & Bigoness, W. J. Race and sex as determinants of ratings by potential employers in a simulated work sample task. Journal of Applied Psychology, 1974, 59, 705-711.
- Hinrichs, J. R. Comparison of "real life" assessments of management potential with situational exercises, paper-and-pencil ability tests, and personality inventories. Journal of Applied Psychology, 1969, 53, 425-432.
- Hinrichs, J. R. Ability correlates in learning a psychomotor task. Journal of Applied Psychology, 1970, 54, 56-64.
- Hirshfeld, S. F. Procedures for Validating Skill Qualification Tests. Arlington, VA.: U.S. Army Research Institute for Behavioral and Social Sciences, 1976. (ERIC Document Reproduction Service No. ED 150 164)
- Holley, W. H., Faid, H. S., & Barnett, N. J. Analyzing performance appraisal systems: An empirical study. Personnel Journal, 1976, 55, 457-459; 463.
- Huck, J. R. Determinants of assessment center ratings for white and Black females and the relationship of these dimensions to subsequent performance effectiveness (Doctoral dissertation, Wayne State University, 1974).

Dissertation Abstracts International, 1975, 35, 3634B. (University Microfilms No. 74-29, 814)

Hunter, D. R. Development of an enlisted psychomotor/perceptual test battery. Catalog of Selected Documents in Psychology, 1976, 6, 62-63, MS 1279.

Hunter, D. R., Maurelle, V. A., & Thompson, N. A. Validation of a psychomotor/perceptual test battery. Catalog of Selected Documents in Psychology, 1977, 7, 122, MS 1616.

Hunter, D. R., & Thompson, N. A. Pilot selection system development. Catalog of Selected Documents in Psychology, 1978, 8, 102, MS 1790.

Hunter, J. E., & Schmidt, F. L. Critical analysis of the statistical and ethical implications of various definitions of Test Bias. Psychological Bulletin, 1976, 83, 1053-1071.

Hunter, J. E., & Schmidt, F. L. Differential and single-group validity of employment test by race: A critical analysis of three recent studies. Journal of Applied Psychology, 1978, 63, 1-11.

Inskip, G. C. Use of psychomotor tests to select sewing machine operators--some negative findings. Personnel Psychology, 1971, 24, 707-714.

Jones, A., & Whittaker, P. A validation technique for performance tests. Occupational Psychology, 1973, 47, 189-192.

Kavanagh, M. J. The content issue in professional appraisal: A review. Personnel Psychology, 1971, 24, 653-668.

Kelleher, C. H., & Quirk, D. A. Age, functional capacity, and work: An annotated bibliography. Industrial Gerontology, 1976, 3, 147-155.

Kelly, E. L. The place of situational testing in evaluating clinical psychologists. Personnel Psychology, 1954, 7, 484-492.

Kirchner, W. K., & Dunnette, M. D. Attitudes toward older workers. Personnel Psychology, 1954, 7, 257-265.

Kuriansky, J. B., Gurland, B. J., Fleiss, J. L., & Cowan, D. The assessment of self-care capacity in geriatric psychiatric patients by objective and subjective methods. Journal of Clinical Psychology, 1976, 32, 95-102.

Levine, E. L., & Flory, A. The in-basket technique: An aid in selection of secretaries. Catalog of Selected Documents in Psychology, 1973, 3, 66, MS 382.

Linn, R. L. Fair test use in selection. Review of Educational Research, 1973, 43, 139-161.

Linn, R. L., & Werts, C. E. Considerations for studies of test bias. Journal of Educational Measurement, 1971, 8, 1-4.

- McClelland, D. C. Testing for competence rather than intelligence. American Psychologist, 1973, 28, 1-14.
- McCormick, E. J., Finn, R. H., & Scheips, G. D. Patterns of job requirements. Journal of Applied Psychology, 1957, 41, 358-364.
- McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 1972, 56, 347-368.
- McGuire, C. Issues in the Assessment of Continuing Professional Competence. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979. (ERIC Document Reproduction Service No. ED 171 788)
- McLaurin, W. A. Validation of a battery of performance tests for prediction of aerospace ground equipment course grades. Catalog of Selected Documents in Psychology, 1974, 4, 62-63, MS 654.
- Meyer, H. H. The validity of the in-basket test as a measure of managerial performance. Personnel Psychology, 1970, 23, 297-307.
- Morse, D. T., & Morse, L. W. A Model for Assessing Effects of Departure from Reality in Performance Testing. Tallahassee, Fla.: Florida State Univ., Career Education Center, 1976. (ERIC Document Reproduction Service No. ED 123 260)
- Mount, M. K., Muchinsky, P. M., & Hanser, L. M. The predictive validity of a work sample: A lab. study. Personnel Psychology, 1977, 30, 637-645.
- Muchinsky, P. M. Utility of work samples in complying with EEOC guidelines. Personnel Journal, 1975, 54(4), 218-226.
- Newsom, R. S., Schulz, J. V., & Friedman, R. B. Evaluation of complex human performance: The promise of computer-based simulation. Evaluation Quarterly, 1978, 2, 421-433.
- Nunnally, J. C., & Wilson, W. H. Method and theory for developing measures in evaluation research. In Struening, E. L., & Guttentag, M. Handbook of Evaluation Research, Vol. I. Beverly Hills, CA.: Sage Publications, 1975. Pp. 227-288.
- O'Connor, P. et al. Improving Reliability in Assessment of Technic Products. Bethesda, MD.: NIH, Bureau of Health Manpower Education, 1974. (ERIC Document Reproduction Service No. ED 094 006)
- O'Leary, L. R. Fair employment, sound psychometric practice, and reality. American Psychologist, 1973, 28, 147-150.
- Olmstead, J. A. et al. Research on Utilization of Assessment Results and Methods. Arlington, VA.: U.S. Army Research Institute for the Behavioral and Social Sciences, 1974. (ERIC Document Reproduction Service No. ED 128 439)

- Osborn, W. C. Essential Dimensions of Performance Tests. Alexandria, VA.: Human Resources Research Organization (HumRRO) Professional Paper No. 1-77, 1977. (ERIC Document Reproduction Service No. ED 137 372)
- Osborn, W. C. Process vs Product Measures in Performance Testing. Alexandria, VA.: Human Resources Research Organization (HumRRO) Professional Paper No. 16-74, 1974.
- Osborn, W. C., & Ford, J. P. Research Methods of Synthetic Performance Tests. Arlington, VA.: U. S. Army Research Institute for the Behavioral and Social Sciences, 1976. (ERIC Document Reproduction Service No. ED 170 346)
- Pederson, K. R. The reliability of a behaviorally recorded work sample (Doctoral dissertation, University of Houston, 1977). Dissertation Abstracts International, 38, 3450B-3451B. (University Microfilm No. 77-29, 682)
- Prien, E. P. The function of job analysis in content validation. Personnel Psychology, 1977, 30, 167-174.
- Prien, E. P., & Ronan, W. W. Job analysis: A review of the research findings. Personnel Psychology, 1971, 24, 371-396.
- Rampton, G. M. A strategy for task analysis and criterion definition based on multidimensional scaling. In Proceedings of the Annual Conference of the Military Testing Association, Vols. I & II. Washington, D.C.: U.S. Coast Guard Institute, 1978. (ERIC Document Reproduction Service No. ED 171 792)
- Robertson, I., & Downs, S. Learning and the predictability of performance. Development of trainability testing in the United Kingdom. Journal of Applied Psychology, 1979, 64, 42-50.
- Ronan, W. W., Anderson, C. L., & Talbert, T. L. A psychometric approach to job performance: Fire fighters. Public Personnel Management, 1976, 5(6), 409-413; 416-422.
- Ryans, D. G., & Frederikson, N. Performance tests of educational achievement. In Lindquist, E. F. (Ed.) Educational Measurement. Washington, D.C.: American Council on Education, 1951. Pp. 455-494.
- Salvendy, G. Selection of industrial operators: The one-hole test. International Journal of Production Research, 1975, 13, 303-321.
- Sanders, J. R. Measurement Problems and Issues Related to Applied Performance Testing. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1976. (ERIC Document Reproduction Service No. ED 124 585)
- Sanders, J. R., & Sachse, T. P. Applied performance testing in the classroom. Journal of Research and Development in Education, 1977, 10(3), 92-104.
- Schmidt, F. L., Berner, J. G., & Hunter, J. E. Racial differences in validity of employment tests: Reality or illusion? Journal of Applied Psychology, 1973, 58, 5-9.

- Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Berner, J. G., & Seaton, F. W. Job sample versus paper-and-pencil trades and technical tests: Adverse impact and examination attitudes. Personnel Psychology, 1977, 30, 187-197.
- Schmidt, F. L., & Hunter, J. E. Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. American Psychologist, 1974, 29, 1-8.
- Schmidt, F. L., Hunter, J. E., & Urry, V. W. Statistical power in criterion-related validation studies. Journal of Applied Psychology, 1976, 61, 473-485.
- Schmidt, F. L. et al. Performance Measurement Feasibility Study: Implications for Manpower Policy. Washington, D.C.: Department of Labor, Manpower Administration, Office of Research, 1974. (ERIC Document Reproduction Service No. ED 128 421)
- Schmitt, N., Mellon, P., & Bylenga, C. Sex differences in validity for academic and employment criteria, and different types of predictors. Journal of Applied Psychology, 1978, 63, 145-150.
- Schwab, D. P., & Heneman, H. G., III. Age stereotyping in performance appraisal. Journal of Applied Psychology, 1978, 63, 573-578.
- Shipley, B. D., jr. Complexity of flight path data as an index of skill in evaluating performances from a flight simulator based job sample test. In Proceedings of the Annual Conference of the Military Testing Association, Vols. I & II. Oklahoma City: U.S. Coast Guard Institute, 1978. Pp. 1193-1204. (ERIC Document Reproduction Service No. ED 171 792)
- Siegel, A. I., & Bergman, B. A. A job learning approach to performance prediction. Personnel Psychology, 1975, 28, 325-339.
- Siegel, A. I., & Bergman, B. A. Nonverbal and Culture Fair Performance Prediction Procedures. I. Background, Test Development, and Initial Results. Washington, D.C.: U.S. Office of Naval Research, Psychological Services, 1972. (ERIC Document Reproduction Service No. ED 069 726)
- Siegel, A. I., Bergman, B. A., & Lambert, J. Nonverbal and culture fair performance prediction. Catalog of Selected Documents in Psychology, 1978, 8, 51, MS 1704.
- Siegel, A. I., Bergman, B. A., & Lambert, J. Nonverbal and Culture Fair Performance Prediction Procedures. II. Initial Validation. Wayne, PA.: Applied Psychological Services, Inc., 1973.
- Siegel, A. I., Fischel, M. A., & Pfeiffer, M. G. Development of performance evaluation measures. Personnel psychophysics: Terminal threshold and signal detection theoretic applications to performance testing. Catalog of Selected Documents in Psychology, 1977, 7, 88, MS 1554.
- Siegel, A. I., & Smith, R. Multidimensional scaling analysis of the job of

- civil defense director. Catalog of Selected Documents in Psychology, 1976, 6, 127.
- Skakun, E. N. et al. A preliminary investigation of computerized patient management problems in relation to other examinations. Educational and Psychological Measurement, 1979, 39, 303-310.
- Smith, M. C., & Downs, S. Trainability assessment for apprentice selection in shipbuilding. Journal of Occupational Psychology, 1975, 48, 39-43.
- Steel, M., Balinsky, B., & Lang, H. A study on the use of a work sample. Journal of Applied Psychology, 1945, 29, 14-21.
- Stodden, R. A., Casals, J., & Schwartz, S. E. Work evaluation and mental retardation. Mental Retardation, 1977, 15(4), 25-28.
- Talbert, T. L., Carroll, K. I., & Ronan, W. W. Measuring clerical job performance. Personnel Journal, 1976, 55, 573-575.
- Thorndike, R. L. Concepts of culture fairness. Journal of Educational Measurement, 1971, 8, 63-70.
- Thornton, G. C., III. The Validity of Assessment Centers. Ft. Collins, CO.: Colorado State Univ., 1971. (ERIC Document Reproduction Service No. ED 050 411)
- Wernimont, P. F., & Campbell, J. P. Signs, samples, and criteria. Journal of Applied Psychology, 1968, 52, 372-376.
- Wiltshire, R. The U.S. Army breakthrough in performance testing. Educational Technology, 1978, 18(3) 35-40.
- Wollack, S. Content validity: Its legal and psychometric basis. Public Personnel Management, 1976, 5, 397-408.